

A scientometrics law about co-authors and their ranking. The co-author core.

M. Ausloos^{1,*}

¹*Beauvallon, r. Belle Jardiniere, 483,
B-4031 Liege, Wallonia-Brussels Federation*

previously at

*GRAPES@SUPRATECS, ULG, B5a Sart-Tilman, B-4000 Liège, Euroland
e-mail address: marcel.ausloos@ulg.ac.be*

(Dated: today)

Let a co-author C have written J (joint) publications with someone. Rank all the co-authors of that individual according to their number of joint publications, giving a rank r to each co-author, starting with $r = 1$ for the most prolific. Examining a finite set of researchers from a group of researchers from statistical physics, over a "long" time interval, a very simple relationship is empirically found between the number of joint publications J by coauthors and their rank of importance, i.e. $J \propto 1/r$. Thus, in the same spirit as for the Hirsch core, one can define a "co-author core", and introduce indices operating on an author. Numerical results adapted to the finite set hereby considered can be meaningfully interpreted. Therefore, variants and generalizations could be later produced in order to quantify co-author roles in a temporary or long lasting stable team.

PACS numbers:

I. INTRODUCTION

In 1926, Lotka discovered that the number of authors n_r is related to the number of published papers r , i.e. [23],

$$n_r \sim n_1/r^2. \quad (1)$$

where n_1 is the number of papers published by the most prolific author. Several other so called *laws* have been predicted or discovered about relations between time, number of publications, number of authors, number of citations, funds, dissertation production, citations, or the number of journals or scientific books.... etc. [7, 11, 20, 30–32, 37, 38]. Scientometrics has become a scientific field in itself [3, 5, 8, 29]. Thus, statistical approaches and models based on the laws and distributions of Lotka, Pareto, Zipf-Mandelbrot, Bradford, Yule, and others, - see Table I for a summary, do provide much useful information for the analysis of the evolution of scientific systems in which development is closely connected to a process of idea diffusion and work collaboration.

More recently, an index, the h -index, has been proposed in order to quantify an individual's scientific research output [17]. A scientist has some index h if h of his/her papers have at least h citations each. *A priori* this h -value is based on journal articles. However, books, monographs, translations, edited proceedings, ... can be included in the measure. The latter may depend on the precision of the examined data basis. No need to say that the best should start from the official publication list of an author. However the number of citations varies according to the data basis. It is rather unusual that an author records by himself the citations of his/her papers. Sometimes, several citations go also unnoticed. The notion of core *for a paper* is also defined, as being a paper which has more than h citations.

A review focusing on the many variants of the h -index, e.g. the a - and e - indices, their computation and standardization can be found in ref. [1]. These indices operate as if it was the single-paper level. However, it is often discussed that inconsistencies may arise from self-citations and multi-authored papers. It is clear, without going into a long discussion, that the role and the impact of co-authors are difficult to measure or even estimate. One may even ask whether there are too many co-authors [25]. Yet, one is often confronted to such questions. To serve as a perspective view point, several considerations from the common literature are briefly outlined here below, without arguing on the pro's and con's.

Several disturbing, or controversial, effects of multi-authorship on citation impact, for example, have been shown in bibliometric studies by Persson et al. in 2004 [28]. However, Glänzel and Thijs,[12] have shown that multi-

*Electronic address: marcel.ausloos@ulg.ac.be

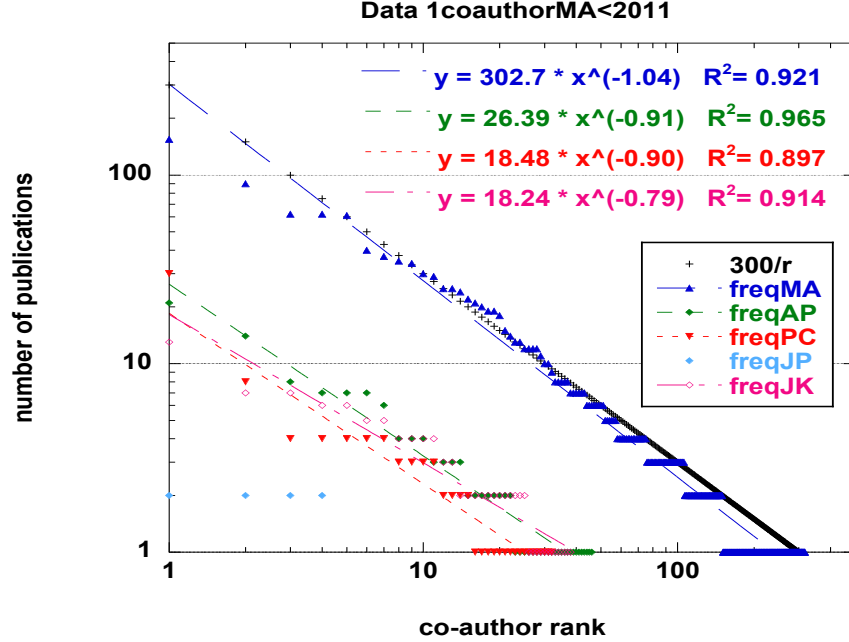


FIG. 1: log-log plot of the number of (joint) publications of coauthors ranked according to rank importance for the 5 team members; a few power law lines are indicated; the $J \simeq 300/r$ law is given as a guide to the eye

authorship does not result in any exaggerate extent of *self-citations*. Moreover, self-citations can indicate some author creativity, or versatility at changing his/her field of research [2, 14–16]

To take into account the effect of multiple *co-authorship* through the h -index, Hirsch [18] even proposed the \hbar index as being the number of papers of an individual that have a citation count *larger than or equal to* the h -index of *all co-authors* of each paper. Of course, $\hbar \leq h$. With the original h -index a multiple-author paper in general belongs to the h -core of some of its coauthors and not belong to the h -core of the remaining coauthors. The \hbar -index, unlike the h -index, uniquely characterizes a paper as belonging or not belonging to the \hbar -core of its authors. However, these considerations emphasize "papers" rather than "authors". Indeed, one focusses on a paper-core, not on an co-author-core.

One may also wonder if co-authors must have all the same "value" in quantifying the "impact" of a paper. Sekercioglu proposed that the k -th ranked co-author be considered to contribute $1/k$ as much as the first author [36], highlighting an earlier proposal by Hagen [13]. At the same time, Schreiber proposed the hm -index [34, 35], counting the papers equally fractionally according to the number of authors; see also Egghe [9] giving an author of an m -authored paper only a credit of c/m if the paper received c citations. Carbone [6] recently also proposed to give a weight m_i^μ to each i -th paper of the j -th individual according to the number m_i of co-authors of this i -th paper, - μ being a parameter at first. Carbone argued that ambiguities in the e.g. h -index distribution of scientist populations are resolved if $\mu \simeq 1/2$. Other considerations can be summarized : (i) Zhang [39] has argued against Sekercioglu hyperbolic weight distribution, as missing the corresponding author, often the research leader. Zhang proposed that weighted citation numbers, calculated by multiplying regular citations by weight coefficients, remain the same as regular citations for the first and corresponding authors, who can be identical, but decreased *linearly* for authors

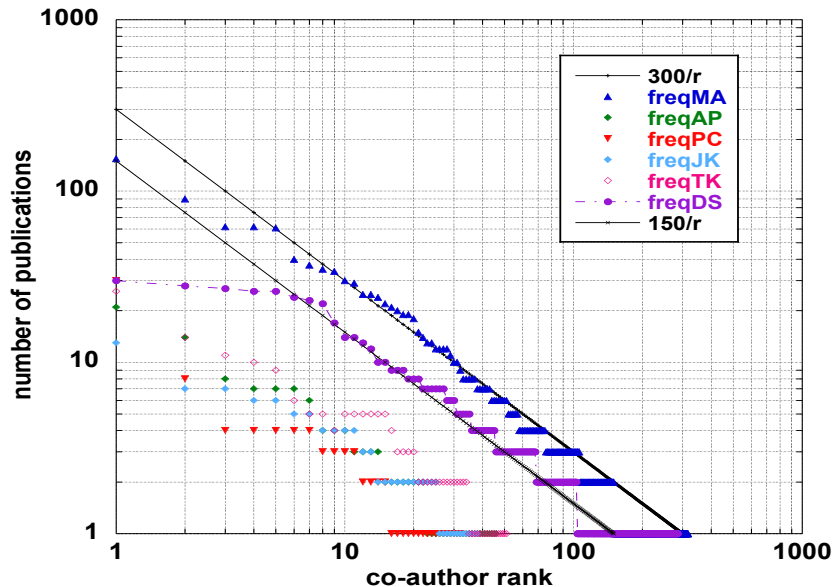


FIG. 2: log-log plot of the number of (joint) publications of co-authors ranked according to importance for the examined team members and outsiders ; (-1) power law lines are indicated for the two most prolific authors, MA and DS. Note the curvature at low rank

with increasing rank; (ii) Galam [10] has recently proposed another fractional allocation scheme for contributions to a paper, imposing in contrast to Zhang, that the total weight of a paper equals 1, *in fine* leading to a *gh*-index favorizing a "more equal" distribution of "co-author's weight" for more frequently quoted papers. Note that it differs from the *hg*-index; see [1].

Other considerations have been given to co-authorship "problems". E.g., Nascimento et al. [26] found out that co-authorship is a small world network. from such a point of view, Börner et al. [4] used a *weighted graph* representation to illustrate the number of publications and their citations. However, even since Newman [27] or more recently Mali et al. [24] and the subsequent works here above recalled, it seems that there have been considerations on the number of co-authors and their "rank", for *one* paper among many others of an individual, but no consideration in the sense of Hirsch, about "*ranking*" over a whole process. The present paper is an attempt to quantify the importance of co-authors, whence co-workers, in scientific publications, over a "long" time interval, thus to suggest further investigations about their effect on/in a team and more generally in a scientific career.

The paper reports on two such aspects. First, in Sect. II, an apparently not reported "law" is presented. Examining a finite set of researchers from a group of researchers, well known to the writer, performing and producing papers in statistical physics, a very simple relationship is empirically found: the number of joint publications J by co-authors C of a researcher and their rank r are related by $J \propto 1/r$, like the number of publications P and their rank, in the *h*-index. The tail of the distribution seems undubious. A deviation occurs for individuals having few co-authors or a limited number of publications. Instead of a -1 slope on a log-log plot, one can observe a possible *Zipf - Mandelbrot* behaviour at small r , i.e.

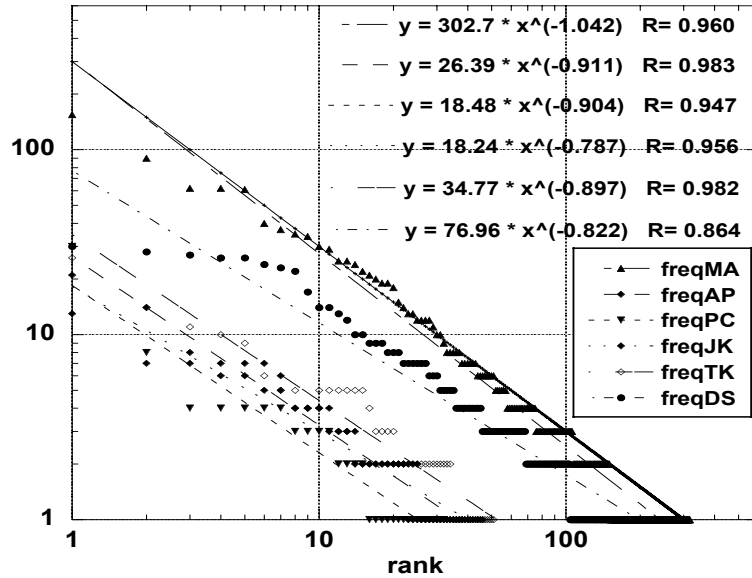


FIG. 3: log-log plot of the number of (joint) publications of co-authors ranked according to importance; the best power law lines are indicated in each case

$$J \propto \frac{1}{(n+r)^\zeta}, \quad (2)$$

with $\zeta \simeq 1$. Next, in the spirit of Hirsch, one can define the core of co-authors of an *individual*, - not of a paper. In parallel to the a_h - or e_h - indices in the publication/citation ranking, one can thus imagine to introduce something called here the m_a - and the $m_a a^2$ - indices, in Sect. III. Numerical illustrations are provided. A short conclusion is found in Sect. IV.

II. DATA

To illustrate the empirical findings, about a relationship between the number of co-authors who have written P publications with someone over some time interval, consider the set of publications produced in statistical mechanics by a group of researchers connected at some time or another with the SUPRATECS Center of Excellence at the University of Liege, Liege, Belgium, at the end of the 20-th century. Let us consider the set made of 5 authors (MA,PC,AP,JP,JK) having various scientific careers, age, expertise or reputation, as given e.g. by their h -index, with other relevant data given in Table II. The CV and list of publications of the team members are available from the author. Let it be emphasized that the joint publications have covered different time spans.

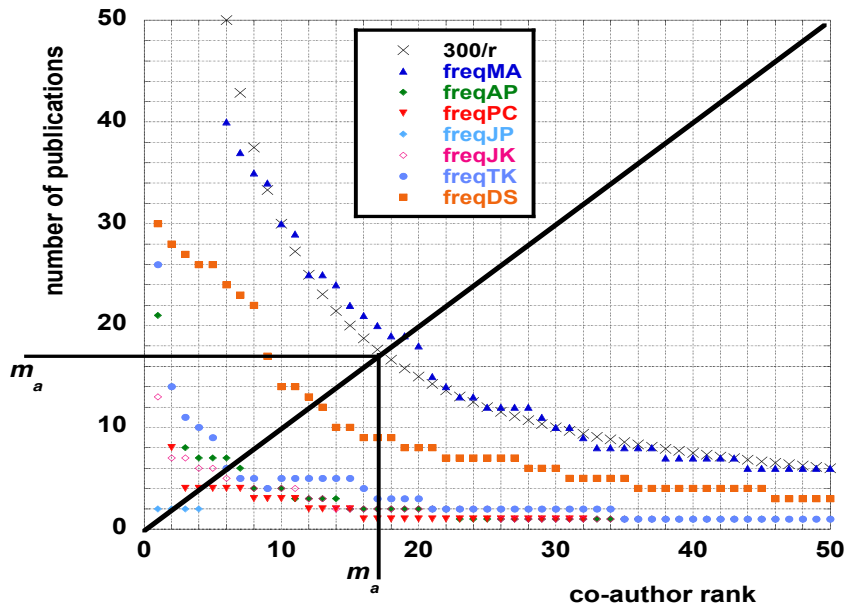


FIG. 4: Selected lin-lin plot of the number of (joint) publications of coauthors ranked according to importance; this allows to define the core of co-authors for the author through an index m_a ; values are given in Table II

Two other scientists are included for comparison: (i) one TK, a younger researcher having collaborated with the group, outside statistical mechanics research; (ii) DS, a researcher known as a guru in the field; see last two columns of Table II.

When the data is not available from the CV or from the co-authors, the data has been taken using Google Scholar search engine. Care has been taken about the correctness of the references and citations. For example, JP has a homonym in another field. The number of citations, leading to the h -index value, includes books when they are recorded as papers in the search engines, papers deposited on arXiv and papers published in proceedings, be they in a journal special issue or in a specific book-like form. Also recall that the number of citations till h divided by h is equal to the a -index. Due to the rather long publication list of MA and AP the total citation count till the end of the examined time interval, i.e., 2010, has not been possible for these. Note that it is somewhat amazing that for such a small number of authors, a hyperbolic Lotka-like law is verified with a $R^2 \simeq 0.995$, though the exponent is close to 3.0 (graph not shown).

More interestingly, a log-log plot of the number (*N.B. By an abuse of language, the data points are called freq for frequency. However no scaling has been made with respect to the total number of publications of each author*) of joint publications between the five team members with either team or other partners, ranked in a decreasing order of joint contributions, is given in Fig. 1. It is remarkable that the power law exponent tends to -1 , the more so if the number of publications of the authors becomes large.

Comparison with the "outsiders" of the main team can be made as a test of the scientific field (ir-)relevance. A log-log plot of the number of joint publications versus ranked co-authors, be they partners or not, ranked in a decreasing order of joint contributions, is given in Fig. 2. It is remarkable that the power law exponent is *very* close to -1 for the

Lotka-Pareto law	gives the number distribution/probability of scientists as a function of the number of papers they wrote	$n_r \sim n_1/r^{1+\alpha}.$ $\alpha \sim 1; n_x = C/(1+x)^{1+\alpha}$ $p(x) = \frac{\alpha}{x_0} \left(\frac{x_0}{x}\right)^{\alpha+1}$
Yule distribution	asymptotically corresponds to Lotka law	$p(x) = \frac{\mu}{\lambda} B\left(x, \frac{\mu}{\lambda} + 1\right) = \alpha B(x, \alpha + 1),$ where $B(x, \alpha + 1) = \Gamma(x)\Gamma(\alpha x + 1)/\Gamma(x + \alpha + 1)$, <i>i.e.</i> , $p(x) \propto \Gamma(\alpha + 1)\alpha/x^{1+\alpha}$
Zipf-Mandelbrot law	ranks scientists by the number of papers they wrote	$C = n_1, \text{ one has } x_r = n_1/(r + a)$ $x_r = \left(\frac{A}{r+B}\right)^\gamma$ $A = (C/\alpha)^{1/\alpha}; \quad B = C/(\alpha k_{\max}^\alpha); \quad \gamma = 1/\alpha.$
Bradford law	reflects the fact that most of the productivity $R(n)$ of relevant articles by scientists are concentrated in a small number n of journals	$R(n) = n_1 \ln\left(\frac{n}{a} + 1\right).$

TABLE I: Bibliometric laws with a few words on their origin and/or usefulness; for more details see Sect. 6.1-6.2 in [37]

most prolific authors, but the curvature at "low rank" indicates that a Zipf-Mandelbrot-like form, Eq. (2) would be more appropriate. This is a very general feature of almost all Zipf plots. However another interesting feature concerns MA for which the $1 < r < 5$ coauthors have an increasing relevance. One way to interpret this feature can be deduced from Table II. It can be observed that the tenure year markedly differs for both authors. It can be understood that DS had more quickly possibilities of collaborations with co-authors than MA who had to list co-authors of hierarchical importance on joint publications during a longer time. This feature is similar to that in the analysis of texts when articles are a mandatory part of the language [22].

Thereafter, one of the co-authors, JP, has been removed from the plots for clarity; JP has in fact a peculiar characteristics, since the researcher has no Ph.D. and has not continued publishing after participating in the team activities. Therefore, only the best fits by a power law to the 4 main team members and the 2 outsiders number of publications versus co-author rank are shown in Fig. 3. The power law is remarkable with deviations from the -1 slope as explained here above.

III. CO-AUTHOR CORE

Similarly to the definition of the "Hirsch core", along the h -index, or also the \hbar -index, concept, one can define the *core of coauthors for an author*. This value, called m_a , is easily obtained from Fig. 4 in the cases so examined through a simple geometrical construction. Similarly to the a -index, one can define the m_a a^2 -index which measures the surface below the empirical data of the number of publications till rank m_a . In so doing an a_a -index could be defined in parallel to some a_h -index. The results are given in Table II.

The interpretation of such results indicates the relative importance of working in a team or not. But also points toward further studies on time of activities. Indeed compare the a_a values for PC and TK, and observe their relative scientific career output as co-authors. Be aware that TK is an experimentalist and PC a theoretician, and started their career at different times. Yet they have a similar record of publications. However, TK, though being associated

	MA	PC	AP	JP	JK	TK	DS
born in	1943	1945	1937	1939	1939	1972	1943
Ph.D. in	1973	1973	37	none	1973	2001.	1970
tenure in	1986	1976	1980	none	1995	2007.	1977
1st publication in	1971	1974	1966	1983	1967	1997	1967
latest recorded publication in	2010	2010	2010	1983	1999	2010	2010
h -index	35	11	10	2	10	6	55
a -index	31.8	26.9	22.4	7	10.	11.	26.
Numb. publications (<2011)	571	34	111	2	60	38	640.
Numb. ed. books (<2011)	9	-	8	-	(2)	-	10.
Most often cited paper numb.cit.	152	127	37	7	537	41	1430
Tot. Numb. citations till h	1113	296	224	14	745	100	8148
Numb. co-authors	317	32	46	4	38	51	285
Numb. Publ. with "best" co-author	155	30	21	2	13	26	30
$\sum_j J_{ij}$	1551	95	134	8	108	181	793
skewness	7.35	4.66	3.18	-	2.18	3.39	3.98
m_a -index	19	4	7	2	5	6	12
$m_a a^2$ -index	810	46	170	4	39	76	264
a_a -index	42.6	11.5	24.3	2	7.8	12.7	22

TABLE II: Data deduced from CV or Google Scholar on hereby examined scientist set

in a loose way with a team, has almost the same a_a (slightly greater than 1) as a stable senior partner, PC. Even though PC has many less co-authors. Similarly, compare a_a for AP and DS, both with a_a above 20, even though their number of co-authors is markedly different, - a ratio = 6.2, corresponding to an equivalent ratio of publications.

IV. CONCLUSION

Two main findings must be outlined as a summary and conclusion. It might have been thought that the number of co-authors of papers over a career might be related to the number of joint publications. But it was not obvious that a simple relationship should be found. In so finding, an interesting new measure of research teams follows.

First, examining a finite set of researchers from a well known active group having performed an activity over decades in statistical physics, a very simple, though unexpectedly simple, relationship has been empirically found between the number of joint publications (J) by coauthors and their rank (r), i.e.

$$J \propto 1/r. \quad (3)$$

Next, in the same spirit as for the Hirsch core, one can define a "co-author core", and introduce indices, like m_a and a_a , operating on an author. Numerical results adapted to the finite set hereby considered can be meaningfully interpreted. Therefore, variants and generalizations could be later produced in order to quantify co-author roles in a temporary team. The finite size of the sample is apparently irrelevant as an argument against the findings. Nevertheless, one could develop the above considerations, through a kind of network study.

As a final point, let it be emphasized that even though co-authorship can be abusive [21], it should not be stupidly scorned upon. Indeed in some cases, co-authorship and output are positively related. For instance, it has been shown that, for economists, more co-authorship is associated with higher quality, greater length, and greater frequency of publications [19, 33]. Yet bibliometric indicators, as those nowadays discussed, can be useful parameters to evaluate the output of scientific research and to give some information on how scientists actually work and collaborate. Of course the present findings and the proposed indices are only a few of the possible quantitative ways to tackle the co-authorship problem. Different other methods can be investigated, with variants as those recalled in the Introduction.

However, they will never be the whole answer to evaluate the career of an individual nor to fund his/her research and team. But they are easy smoke screens.

Acknowledgment

The authors gratefully acknowledges stimulating and challenging discussions with many wonderful colleagues at several meetings of the COST Action MP-0801, 'Physics of Competition and Conflict'. In particular thanks to O. Yordanov for organising the May 2012 meeting "Evaluating Science: Modern Scientometric Methods", in Sofia, and challenging the author to present new results. All scientific colleagues mentioned in the text have frankly commented upon the manuscript and enhanced its content .

-
- [1] Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics* 3, 273-289.
 - [2] Ausloos, M., Lambiotte, R., Scharnhorst, IA., Hellsten, I. (2008). Andrzej Pekalski networks of scientific interests with internal degrees of freedom through self-citation analysis, *Int. J. Mod. Phys. C* 19, 371-384.
 - [3] Beck, I. M. (1984). A method of measurement of scientific production. *Science of Science* 4, 183 – 195.
 - [4] Börner, K., Dall'Asta, L., Ke, W., Vespignani, A. (2005) Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams *Complexity* 10, 57-67.
 - [5] Bruckner, E., Ebeling, W., Scharnhorst, A. (1990). The application of evolution models in scientometrics. *Scientometrics* 18, 21 – 41.
 - [6] Carbone, V.(2012). Fractional counting of authorship to quantify scientific research output *arxiv 1106.0114v1*
 - [7] Chung, K.H., Cox, R. A. K. (1990). Patterns of productivity in the finance literature: A study of the bibliometric distributions. *Journal of Finance* 45, 301 – 309.
 - [8] Egghe, L., Rousseau, R. (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science, Elsevier, Amsterdam.
 - [9] Egghe, L. (2008d). Mathematical theory of theh- and g-index in case of fractional counting of authorship. *Journal of the American Society for Information Science and Technology* 59, 1608-1616.
 - [10] Galam, S. (2011). Tailor based allocations for multiple authorship: a fractional gh-index. *Scientometrics* 89, 365-379
 - [11] Gilbert, G.N. (1978). Measuring the growth of science: A review of indicators of scientific growth. *Scientometrics* 1, 9 – 34.
 - [12] Glänzel, W., Thijs, B. (2004). Does co-authorship inflate the share of self-citations?. *Scientometrics* 61, 395-404.
 - [13] Hagen, N. T. (2009) Credit for coauthors. *Science* 323, 583.
 - [14] Hellsten, I., Lambiotte, R., Scharnhorst, A., Ausloos, M. (2006). A journey through the landscape of physics and beyond - the self-citation patterns of Werner Ebeling". in: Irreversible Prozesse und Selbstorganisation. T. Poeschel, H. Malchow and L. Schimansky-Geier, Eds. (Logos Verlag, Berlin, 2006) pp. 375-384
 - [15] Hellsten, I., Lambiotte, R., Scharnhorst, A., Ausloos, M. (2007). Self-citations, co-authorships and keywords: A new method for detecting scientists field mobility? *Scientometrics* 72, 469 - 486.
 - [16] Hellsten, I., Lambiotte, R., Scharnhorst, A., Ausloos, M. (2007). Self-citations networks as traces of scientific careers. In Proceedings of the ISSI 2007, 11th International Conf. of the Intern. Society for Scientometrics and Informetrics, CSIC, Madrid, Spain, June 25-27, 2007. Ed. by D. Torres-Salinas; H. Moed, Vol. 1, pp. 361-367
 - [17] Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences USA* 102, 16569-16572.
 - [18] Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics* 85, 741-754.
 - [19] Hollis, A. (2001). Co-authorship and the output of academic economists, *Labour Economics* 8, 505-530.
 - [20] Kealey, T. (2000). More is less. Economists and governments lag decades behind Derek Price's thinking. *Nature* 405, 279.
 - [21] Kwok, L. S. (2005). The White Bull effect: abusive coauthorship and publication parasitism. *Journal of Medical Ethics* 31, 554-556.
 - [22] Li, W. (2002). Zipf's law everywhere. *Glottometrics* 5, 15 – 41.
 - [23] Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 16, 317 – 324.
 - [24] Mali, F., Kronegger, L., Doreian, P., Ferligoj, A. (2012). Chapter 6, Dynamic Scientific Co-Authorship Networks in Models of Science Dynamics : Encounters Between Complexity Theory and Information Sciences, Andrea Scharnhorst, Katy Börner, and Peter van den Besselaar, Eds. (Springer Verlag Berlin Heidelberg).
 - [25] McDonald, K. A. (1995). Too Many Co-Authors?. *Chronicle of Higher Education* 41, 35-36
 - [26] Nascimento, M. A., Sander, J., Pound, J. (2003). Analysis of SIGMOD's co-authorship graph. *ACM SIGMOD Record Homepage archive* 32, 8 - 10.
 - [27] Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences USA* 101, 5200-5205.
 - [28] Persson, O., Glänzel, W., Danell, R. (2004). Inflationary Bibliometric Values: The Role of Scientific Collaboration and the Need for Relative Indicators in Evaluative Studies. *Scientometrics* 60, 421-432.

- [29] Potter, W. G. (1988). Of Making Many Books There is No End: Bibliometrics and Libraries. *The Journal of Academic Librarianship* 14, 238a-238c.
- [30] Price, D. J. de Solla. (1963). Little Science, Big Science. Columbia University Press, New York.
- [31] Price, D. J. de Solla, Gürsey, S. (1975). Some statistical results for the numbers of authors in the states of the United States and the nations of the world. pp. 26 – 34 in: Who is Publishing in Science, 1975 Annual. Institute for Scientific Information, Philadelphia.
- [32] Price, D. J. de Solla. (1978). Science since Babylon. Yale University Press, New Haven.
- [33] Sauer, R.D. (1988). Estimates of the returns to quality and coauthorship in economic academia. *The Journal of Political Economy* 96, 855-866.
- [34] Schreiber, M. (2008). To share the fame in a fair way, hm for multi-authored manuscripts. *New Journal of Physics* 10 (040201), 1-9.
- [35] Schreiber, M. (2008). A modification of the h-index: The h(m)-index accounts for multi-authored manuscripts. *Journal of Informetrics* 2, 211-216.
- [36] Sekercioglu, C.H. (2008). Quantifying coauthor contributions. *Science* 322, 371.
- [37] Sekercioglu, C.H. (2009). Response from Cagan H. Sekercioglu to Hagen (2009). *Science* 30, 583.
- [37] Vitanov, .K., Ausloos, M. (2012). Knowledge epidemics and population dynamics models for describing idea diffusion, in *Models of Science Dynamics : Encounters Between Complexity Theory and Information Sciences*, Andrea Scharnhorst, Katy Börner, and Peter van den Besselaar, Eds. Springer Verlag Berlin Heidelberg, Ch. 3, pp. 69 - 125.
- [38] Yablonsky, A.I. (1980). On fundamental regularities of the distribution of scientific productivity. *Scientometrics* 2, 3 – 34.
- [39] Zhang, C.T., . (2009). A proposal for calculating weighted citations based on author rank. *EMBO Reports* 10, 416-417.